Stat 160 Fall 2008
Linear Regression and Least Squares

In the box *Least-Squares Regression Line* on page 119, Moore writes:

> The **least-squares regression line** of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

Furthermore, Moore indicates that "one reason for the popularity of the least-squares regression line is that the problem of finding the line has a simple answer."

To explain the first quote from Moore, it is helpful to suppose that we have bivariate data $(x_1, y_1), \ldots, (x_n, y_n)$ and that we are interested in constructing a linear regression model to fit this data. That is, we want to find the equation of a line of the form

$$\hat{y} = a + bx. \qquad (*)$$

Notice that we write y_i for our data points and $\hat{y}_i$ for the corresponding points on the line. Thus, the vertical distance of the data point to the line is simply $y_i - \hat{y}_i$ and so the sum of the squares of the vertical distances is

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \qquad (**)$$

(This is illustrated in Figure 5.2 on page 119.)

At this point, we DO NOT know the values of a and b. Our goal is to find a and b to minimize $(**)$. As noted in class, the solution to this problem typically involves calculus and/or linear algebra. However, an elementary solution involving only high school algebra is possible if we make the following assumption.

The regression line $(*)$ passes through the point $(\overline{x}, \overline{y})$.

Making this assumption, we know that the point $(\overline{x}, \overline{y})$ is a solution to the regression equation $\hat{y} = a + bx$ which means that

$$\overline{y} = a + b\overline{x}.$$

Solving this for a tells us that

$$a = \overline{y} - b\overline{x}.$$

We can now substitute this back into $(*)$ to conclude

$$\hat{y} = a + bx = \overline{y} - b\overline{x} + bx = \overline{y} + b(x - \overline{x}).$$

We now return to our goal of trying to minimize $(**)$. Since we now have an expression for $\hat{y}$ involving only b given by the previous equation, we can substitute into the expression for SSE and conclude

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - \overline{y} - b(x_i - \overline{x}) \right]^2 = \sum_{i=1}^{n} \left[(y_i - \overline{y}) - b(x_i - \overline{x}) \right]^2.$$

In order to manipulate this expression, we begin by expanding the square to find

$$\left[(y_i - \overline{y}) - b(x_i - \overline{x}) \right]^2 = (y_i - \overline{y})^2 - 2b(x_i - \overline{x})(y_i - \overline{y}) + b^2(x_i - \overline{x})^2$$

so that

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2b\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) + b^2\sum_{i=1}^{n}(x_i - \overline{x})^2. \qquad (\dagger)$$

If we now divide both sides by $n-1$, then we have

$$\text{SSE} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 - \frac{2b}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) + \frac{b^2}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Notice that the definition of variance implies

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = s_y^2 \quad \text{and} \quad \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s_x^2.$$

Furthermore, the definition of regression is

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$$

so that

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = s_x s_y r.$$

Thus, we find (†) is equivalent to

$$\text{SSE} = s_y^2 - 2s_x s_y r b + s_x^2 b^2.$$

We now see that this expression for SSE is a quadratic (parabola) in b written in general form. We can now complete the square to write it in standard form. That is,

$$\text{SSE} = s_y^2 - 2s_x s_y r b + s_x^2 b^2 = s_x^2\left[b^2 - 2r\frac{s_y}{s_x}b\right] + s_y^2 = s_x^2\left[b^2 - r\frac{s_y}{s_x}\right]^2 + s_y^2 + r^2 s_y^2.$$

At this point, it is clear that SSE represents a parabola in b which "opens up" and is centred at

$$\left(r\frac{s_y}{s_x}\,,\ s_y^2 + r^2 s_y^2\right).$$

Therefore, the minimal value of SSE is at this centre point, namely $s_y^2 + r^2 s_y^2$, and it occurs when

$$b = r\frac{s_y}{s_x}.$$

In conclusion, if we assume that the regression line passes through the point $(\overline{x}, \overline{y})$, then we find that the equation of the regression line is

$$\hat{y} = a + bx$$

where

$$b = r\frac{s_y}{s_x} \quad \text{and} \quad a = \overline{y} - b\overline{x}.$$